Quantifying transmission dynamics of acute hepatitis C virus infections in men having sex with men using phylodynamics

Gonché Danesh[1◕*], Victor Virlogeux[2◕], Christophe Ramière[3¤], Caroline Charre[3], Laurent Cotte[4‡], Samuel Alizon[1‡]

1 MIVEGEC (UMR CNRS 5290, IRD, UM), Montpellier, France

2 Clinical Research Center, Croix-Rousse Hospital, Hospices Civils de Lyon, France

3 Virology Laboratory, Croix-Rousse Hospital, Hospices Civils de Lyon, France

4 Infectious Diseases Department, Croix-Rousse Hospital, Hospices Civils de Lyon, France

◕These authors contributed equally to this work.

‡These authors also contributed equally to this work.

* Corresponding author: gonche.danesh@ird.fr

Abstract

Opioid substitution and syringes exchange programs have drastically reduced hepatitis C virus (HCV) spread in France but HCV sexual transmission in men having sex with men (MSM) has recently arose as a significant public health concern. The fact that the virus is transmitting in a heterogeneous population, with 'new' and 'classical' hosts, makes prevalence and incidence rates poorly informative. However, additional insights can be gained by analyzing dated phylogenies inferred from dated virus genetic sequence data. Here, using such a phylodynamics approach based on Approximate Bayesian Computation, we estimate key epidemiological parameters of an ongoing HCV epidemic in MSM in Lyon (France). We show that this epidemics in MSM is largely independent from the 'classical' HCV epidemics and that its doubling time is one order of magnitude lower (55.6 days *versus* 511 days). These results have practical implications for HCV control and open new perspective for using virus genomics in public health.

Background

The burden of chronic hepatitis C virus (HCV) infection is currently estimated to 71 million infections worldwide [1, 2]. The virus being exclusively a human agent, the World Health Organisation (WHO) and several countries have issued recommendations towards its 'elimination'. This means the absence of a significant transmission in a given epidemiological context and is defined by a 80% reduction in new chronic infections and a 65% decline in liver mortality by 2030 [2]. HIV-HCV coinfected patients are considered a key population because of the shared routes of transmission between the two viruses [3] and because of the increased severity of chronic HCV infection in coinfections [4–6]. This population was therefore targeted with priority, leading to high treatment uptake in countries with affordable access to treatments. This was complemented by successful harm reductions interventions such as needle-syringes exchange and opiates substitution programs, as well as the high level of enrolment into care of HIV-infected patients. As a result, the prevalence of active HCV infections drastically dropped during the recent years in HIV-HCV coinfected patients in several European countries, increasing the hope that HCV elimination was an attainable goal [7–10].

Unfortunately, sexual transmission of HCV in HIV-infected men having sex with men (MSM) recently arose as a significant phenomenon, driven by unprotected sex and drug use in the context of sex ('chemsex') and by potentially traumatic practices such as fisting and sharing sextoys [11–13]. This is the case in area of Lyon (France), where HCV incidence has been shown to increase concomitantly with a shift in the profile of infected hosts [14]. Understanding and quantifying this recent increase is the main goal of this study.

Several modeling studies have highlighted the difficulties to control HCV infection in HIV-infected MSM in the absence of harm reduction interventions [12, 15]. Furthermore, we recently described the spread of HCV from HIV-infected to HIV-negative MSM, using or not HIV pre-exposure prophylaxis (PrEP), through sharing of high-risk practices between these populations [14]. This resulted in an alarming incidence of acute HCV infection in both HIV-infected and PrEP-using MSM in France in 2016-2017 [13]. Additionally, while PrEP-using MSM are regularly screened for HCV, those who are HIV-negative and do not use PrEP may remain undiagnosed and untreated for years. Since we know little about the population size and the practices of HIV-negative MSM who do not use PrEP, these recent epidemiological events could jeopardize the goal of HCV elimination by creating a large pool of infected and undiagnosed patients, pursuing high-risk practices that could fuel new infections in intersecting populations. Furthermore, the epidemiological dynamics of HCV infection have mostly been studied in intravenous drug users (IDU) [16–19] and in the general population [20, 21]. Results from these populations are not easily transferable to other populations, which calls for a better understanding of the epidemiological characteristics of HCV sexual transmission in MSM.

Given the lack of knowledge about the focal population driving the increase in HCV incidence, we investigated virus sequence data using phylodynamics methods. This research field has been blooming over the last decade [22–24] and hypothesizes that the way rapidly evolving viruses spread leaves 'footprints' in their genomes. By combining epidemiological modeling, statistical analyses and phylogenies of infections, where each leaf corresponds to the virus sequence isolated from a patient, current methods can estimate key transmission parameters of viral epidemics. This

framework has been successfully applied to HCV epidemics [25–28] but the epidemics we study is particularly challenging to analyze because the focal population is heterogeneous, with 'classical' hosts (typically HIV-infected patients with a history of opioid intravenous drug use) and 'new' hosts (HIV-infected and HIV-negative MSM, detected during or shortly after acute HCV phase, potentially using recreational drugs such as cocaine or cathinones). To address this issue, we used a framework based on Approximate Bayesian Computation (ABC) that was recently developed and validated [29]. We implemented an epidemiological model with two host types ('classical' and 'new'), where each infection can generate secondary infections before ending (see the Methods). By analyzing dated virus sequences, we estimated the date of onset of the HCV epidemics in the 'classical' hosts, and in the 'new' hosts, the level of mixing between the hosts types, and, for each host type, the duration of the infectious period and the basic reproduction ratio (i.e. the number of secondary infections, [30]). This allowed us to show that the doubling time of the epidemics in 'new' hosts is dramatically higher than that in 'classical' hosts, therefore emphasising the urgent need for public health action.

Results

The time-scaled phylogeny inferred from the dated virus sequences reveals that 'new' hosts (in red) tend to cluster together (Figure 1). This suggests a high level of assortativity (i.e. each infected host tends to infect hosts from the same type). Furthermore, the estimate for the root of the phylogeny, that is the onset of the epidemics in the studied area, is in the early 1980s, which appears consistent with epidemiological data.

As for any bayesian inference method, we need to assume prior distribution for each parameter (in grey in Figure 2) in order to infer posterior distributions (in red in Figure 2). Priors were voluntarily assumed to be large and uniformly distributed so as to be as little informative as possible. The only exception was the date of onset of the epidemics, for which we used the output of the phylogenetic analysis as a prior. For the date of the second epidemics, we assumed that it took place after 1997 based on epidemiological data. The width of the posterior distribution indicates our ability to infer a parameter.

The ABC phylodynamics approach allows us to go beyond a visual description and to quantify several epidemiological parameters. For instance, we can narrow down the estimation for the origin of the epidemic (in 'classical' hosts only) to 1976[1969; 1980] (numbers in brackets indicate the 95% Highest Posterior Density, or HPD). The epidemic in the second host type is estimated to have started in 2001[1998; 2005].

Regarding the level of assortativity between host types, that is the extent to which a host of a given type interacts with hosts of the same type, we estimate a_1 to be 0.96[0.86; 0.99] and a_2 to be 0.86[0.72; 0.99]. Therefore, hosts appear to preferentially interact with hosts from the same type and this effect seems more pronounced for 'classical' hosts.

The phylodynamics approach also allows us to infer the duration of the infectious period. Here, assuming that this parameter remains constant for a given host type over time, we estimate it to be 1.7 years [0.46; 9.17] for 'classical' hosts and 0.4 years [0.26; 0.65] for 'new' hosts.

The basic reproduction numbers, i.e. the number of secondary infections caused by a given host

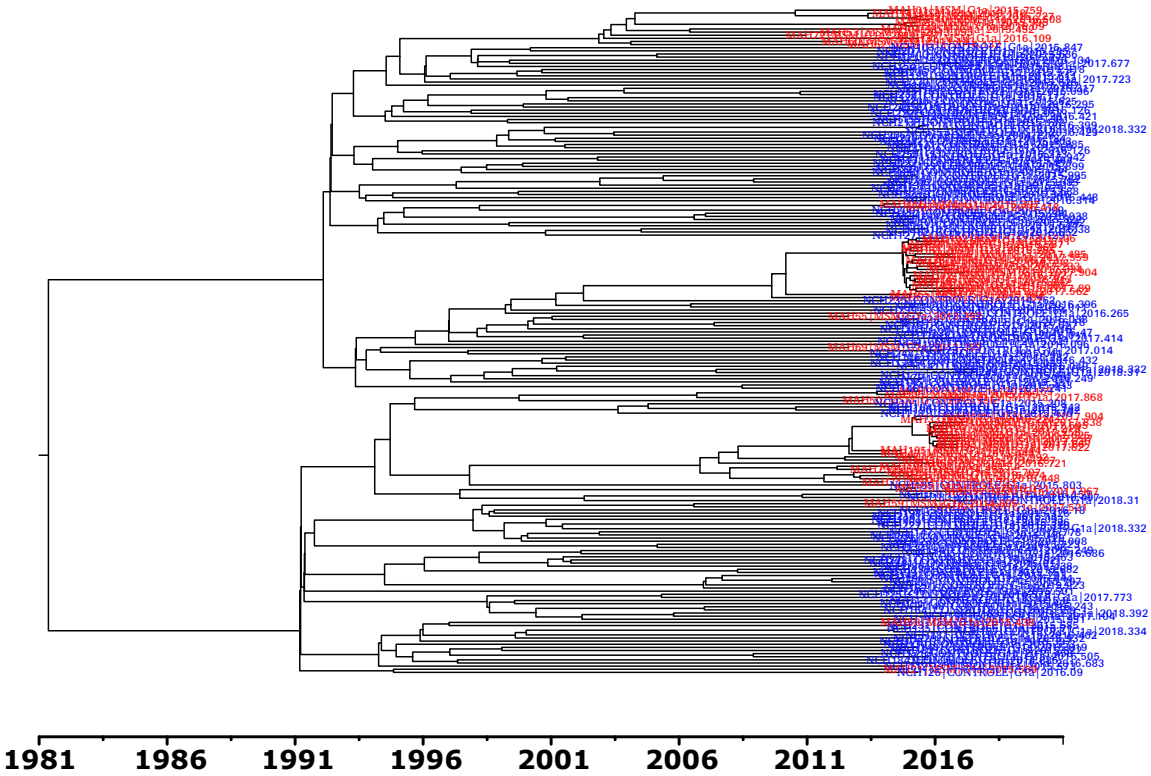

Fig 1. Phylogeny of HCV infections in the area of Lyon (France). 'Classical' hosts are in blue and 'new' hosts are in red. The phylogeny was estimated using maximum-likelihood methods (PhyML) and then rooted in time using bayesian inference (Beast2). See the Methods for additional details.

over the infectious period, was estimated to have decreased from $5.94[3.24; 8.61]$ to $1.80[1.12; 2.48]$ for 'classical' hosts, after the introduction of the third generation HCV test in 1997. We also estimate that 'new' hosts transmit HCV $6.50[2.56; 9.81]$ times more than 'classical' hosts (parameter ν). Using all these inferences, we can calculate the reproduction number in 'new' hosts, $R_0^{(2),t_3}$ (see the Methods), which is $2.35[0.55; 8.05]$.

To better show the differences between the two host types, we compute the epidemics doubling times (t_D), which is the time it takes for an infected population to double in size, for each type of host, assuming a full assortativity. We find that since 1997,the $t_D^{(1),t2}$ could be estimated to 511.0 days ($[0.58; 10.13]$ years) for the 'classical' hosts, whereas the $t_D^{(2),t3}$ was estimated to 55.56 days ($[0; 3.51]$ years) for the 'new' hosts. Before 1997, the $t_D^{(1),t1}$ was estimated to 83 days ($[0.05; 1.60]$ years) for the 'classical' hosts. We show the densities of theses three doubling times in Supplementary Figure S2.

In Supplementary Figure S3, we show the correlations between parameters in the posterior distributions (Figure 3). We mainly find that the R_0 in 'classical' hosts after the introduction of the third generation of HCV detection tests (i.e. $R_0^{(1),t_2}$) is negatively correlated to ν and positively correlated to γ_2. This makes sense because a rapid growth of the epidemic in 'classical' hosts imposes

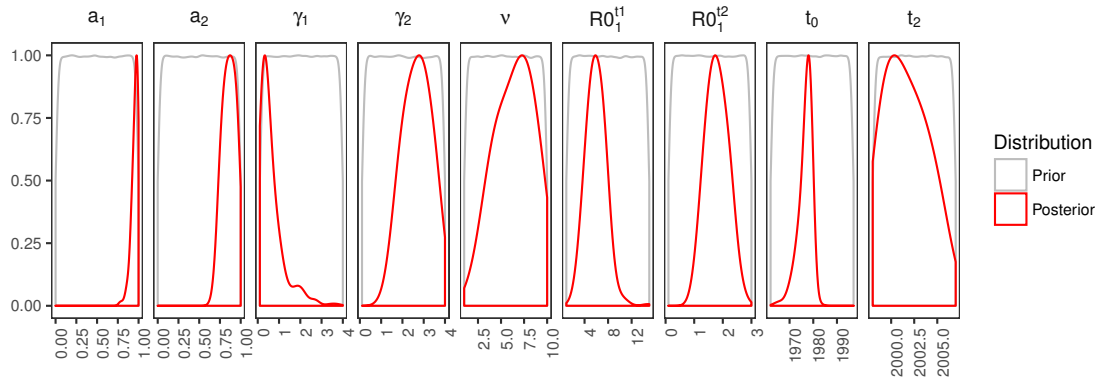

Fig 2. Parameter prior and posterior distributions. Prior distributions are in grey and posterior distributions of the regression-ABC are in red. The thinner the posterior distribution, the more accurate the inference.

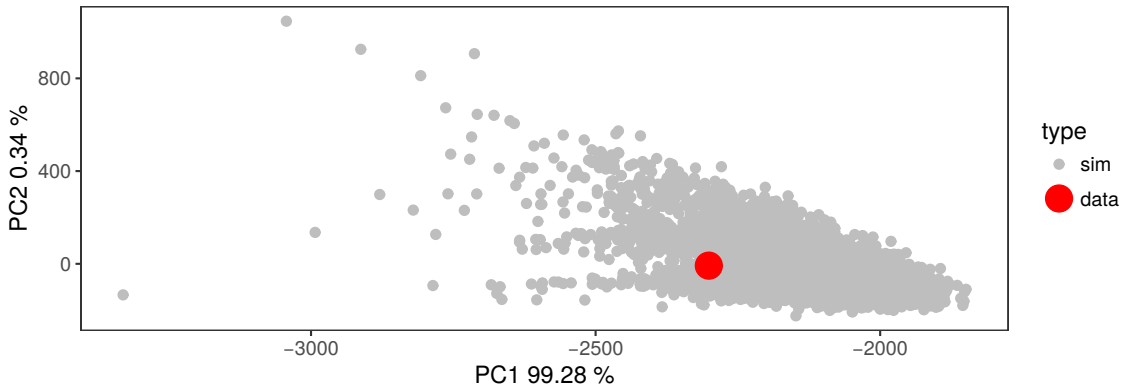

Fig 3. Parameteric bootstrap illustration. PCA graph where each dot represents a vector of summary statistics of a data. The 1,000 simulated data are in grey, and the target data is in red.

a lower growth in 'new' hosts. $R_0^{(1),t_2}$ is also slightly negatively correlated to γ_1, which probably comes from the fact that epidemics with the same R_0 but a longer infection duration have a lower doubling time and therefore a weaker epidemiological impact. Overall, these correlations do not affect our main results, especially the pronounced difference in infection periods (γ_1 and γ_2).

To validate these results, we performed a parametric bootstrap analysis by simulating phylogenies using our posterior distributions to determine whether these are similar to the target dataset (see the Methods). In Figure 3, we see that the target data in red, i.e. the summary statistics from the phylogeny shown in Figure 1, lies in the middle of the phylogenies simulated using the posterior data. Even if we use the 95% HPD of the posterior but assume a uniform distribution instead of the true posterior distribution, we find that the target phylogeny lies outside the cloud of simulations (see Supplementary figure S4). These results confirm that the posterior distributions we infer are highly informative regarding the phylogeny shape.

Finally, to further validate the accuracy of our inference method, we used simulated data to

perform a leave one out cross-validation (see the Methods). As shown in Supplementary Figure S5, the relative error made for each parameter inference is limited and comparable to what was found using a simpler model [29]. Two exceptions are the rate at which 'new' hosts clear the infection (γ_2) and their level of assortativity (a_2). However, this is likely to be due to the constraint imposed by the shape of the target phylogeny itself rather than by the method. In general, this cross-validation goes beyond the scope of this epidemiological model because, for instance, assortativity values can vary between 0 and 1, whereas for the phylogenetic structure we studied with a high degree of clustering, we expect it should be close to 1.

Discussion

Over the last years, an increase in HCV incidence has been witnessed in the area of Lyon (France), which involves both HIV-positive and HIV-negative men having sex with men (MSM) and appears to be driven by sexual transmission. Similar trends have been described in Amsterdam [31] and, more recently, in Switzerland [32]. Achieving a quantitative analysis of this epidemics is required to optimise public health interventions but extremely challenging because the monitoring of the population at risk is limited and because classical tools in quantitative epidemiology such as incidence time series are poorly informative in such a heterogeneous population. To address this issue, we analysed virus sequence data using phylodynamics. In order to account for host heterogeneity, we extended and validated an existing framework relying on Approximate Bayesian Computation [29].

From a public health point of view, these results have two major implications. First, there is a strong assortativity within the 'classical' and 'new' hosts. This can be seen qualitatively from the phylogeny (Figure 1) but the ABC approach allows us to quantify it and to show that assortativity might be higher for the 'classical' hosts. The second strong result has to do with the massive difference in doubling time of the epidemics between the two host types. Indeed, the current spread of the epidemics in 'new' hosts appears to be at least comparable the spread in the early 1990s before the advent of the third generation tests. That the duration of the infectious period in new hosts is in the same order of magnitude as the time until treatment suggests that the majority of the infection may be occurring during the acute phase. This underlines the necessity to act rapidly upon detection, for instance with emphasising the importance of protection measures (condom use) and treatment initiation even during the acute phase [33]. A better understanding of the underlying contact networks therefore seems essential and, with that respect, next generation sequence data could be particularly informative [34–36].

Two legitimate interrogations about the study have to do with the sampling scheme and the assessment of the host type. Regarding the sampling, the proportion of the infections in 'new' that are sampled is estimated to be high. For the 'classical' hosts, we selected a representative subset of the patients detected in the area. Regarding the host type, this was assessed by clinicians independently of the sequence data. The main criterion used was the infection stage (acute or chronic), which was complemented by other epidemiological criteria (history of intravenous drug use, HIV status). In order to test whether the infection stage (acute vs. chronic) might not explain the data as well, we developed an alternative model where all infected hosts first go through the acute

phase before recovering or progressing to the chronic phase. As for the model with two host types, we used 3 time intervals. Interestingly, it was almost impossible to simulate phylogenies under this model. This is most likely due to the fact that there cannot be an assortativity parameter in this alternative model (all new infections must be acute), which makes it more difficult to reproduce the observed phylogeny.

The phylodynamics analysis raised technical challenges because of the known heterogeneity in the host population. To our knowledge, only two studies have recently tackled this issue. In 2018, a group used the structured coalescent model, to investigate the importance of accounting for so-called 'superspreaders' in the recent ebola epidemics in West Africa [37]. The same year, another group used the birth-death model to study the effect of drug resistance mutations on the R_0 of HIV strains [38]. Both of these are implemented in Beast2. However,the birth-death model is unlikely to be directly applicable to our HCV epidemics because it links the two epidemics via mutation (a host of type A becomes a host of type B), whereas in our case the linking is done via transmission (a host of type A infects a host of type B).

This study shows that the ABC approach, which had been validated for simple epidemiological models such as Susceptible-Infected-Recovered [29] can also be applied to more elaborate models that most phylodynamics methods have difficulties to include. Further increasing the level of details in the model will require further analyses. This may require to increase the number of simulations but also to introduce new summary statistics. Another promising perspective would be to combine sequence and incidence data. Although this could not be done here due to the limited sampling, such a combination of different sources of data can be readily performed in an ABC framework.

Material and methods

Epidemiological data

The Dat'AIDS cohort is a collaborative network of 23 French HIV treatment centers covering approximately 25% of HIV-infected patients followed in France (Clinicaltrials.gov ref NCT02898987). The epidemiology of HCV infection in the cohort has been extensively described from 2000 to 2016 [39–41]. The incidence of acute HCV infection has been estimated among HIV-infected MSM between 2012 and 2016 and among HIV-negative MSM enrolled in PrEP between in 2016-2017 [13]. The epidemiology of acute HCV infection, including incidence estimates, in HIV-infected and HIV-negative MSMs has been described from 2014 to 2017 [14].

HCV sequences data

HCV molecular sequences of all MSM patients ($N = 68$) diagnosed with acute HCV genotype 1a infection at the Infectious Disease Department of the Hospices Civils de Lyon, France, and for whom NS5B sequencing was performed between January 2014 and December 2017 were considered. HCV genotype 1a isolated from $N = 145$ non-MSM, HIV-negative, male patients of similar age were analysed by NS5B sequencing at the same time for phylogenetic analysis. This study was conducted in accordance with French ethics regulations. All patients gave their written informed consent to

allow the use of their personal clinical data. The study was approved by the Ethics Committee of Hospices Civils de Lyon.

HCV testing and sequencing

HCV RNA was detected and quantified using the Abbott RealTime HCV assay (Abbott Molecular, Rungis, France). The NS5B fragment of HCV was amplified between nucleotides 8256 and 8644 by RT-PCR as previously described and sequenced using the Sanger method. Electrophoresis and data collection were performed on a GenomeLab^{TM} GeXP Genetic Analyzer (Beckman Coulter). Consensus sequences were assembled and analysed using the GenomeLab^{TM} sequence analysis software. The genotype of each sample was determined by comparing its sequence with HCV reference sequences obtained from GenBank.

Nucleotide accession numbers

All HCV NS5B sequences isolated in MSM and non-MSM patients reported in this study were submitted to the GenBank database. The list of Genbank accession numbers for all sequences is provided in Appendix.

Dated viral phylogeny

We inferred a maximum likelihood phylogeny using PhyML v3.0 software complemented with the Smart Model-Selection (SMS) software, from the ATGC platform [42,43], to perform model selection. The SMS tool selected the GTR+Γ+I model with four substitution rate categories. The maximum likelihood phylogeny was then rooted using BEAST v2.4.8 [44]. To do so, two trees were built using BEAST2 according to two molecular clock models: either relaxed or strict [45]. We performed a model comparison with Tracer v.1.6.0 using the AIC criterion. The strict molecular clock model had a lower AIC value and was therefore considered to be the best model. The date of the last common ancestor was estimated at 1981.34 with a 95% Highest Posterior Density (HPD) of [1962.03; 1997.26].

Epidemiological model and simulations

We assumed a Birth-Death model with two hosts types (Supplementary Figure S1) with 'classical' hosts (numbered 1) and new hosts (numbered 2). This model is described by the following system of ordinary differential equations (ODEs):

$$\frac{dI_1}{dt} = a_1\beta I_1 + (1-a_2)\nu\beta I_2 - \gamma_1 I_1 \tag{1a}$$

$$\frac{dI_2}{dt} = a_2\beta\nu I_2 + (1-a_1)\beta I_1 - \gamma_2 I_2 \tag{1b}$$

In this model, transmission events are possible within each type of hosts and between the two types of hosts at a transmission rate β. The parameter ν corresponds to the transmission differential between the number of partners of the classical hosts (I_1) and that of the new hosts (I_2). Individuals

Table 1. Prior distributions for the birth-death model parameters over the three time intervals. t_0 is the date of origin of the epidemics in the studied area, t_1 is the date of introduction of 3rd generation HCV tests, t_2 is the date of emergence of the epidemic in 'new' hosts and t_f is the time of the most recent sampled sequence.

Interval	γ_1	γ_2	ν	$R_0^{(1)}$	a_i
$[t_0, t_1]$	Unif$(0.1, 4)$	Unif$(0.1, 4)$	0	Unif$(0.9, 15)$	Unif$(0, 1)$
$[t_1, t_2]$				Unif(0.13)	
$[t_2, t_3]$			Unif$(1, 10)$		

I_i can be 'removed' from the infectious compartment i via infection clearance, host death or change in host behaviour (e.g. condom use). This event occurs at a removal rate γ_1. The assortativity between host types is given by the a_i (a value close to 1 means there is very little transmission to hosts from the other type).

The basic reproduction number (denoted R_0) is the number of secondary cases caused by an infectious individual in a fully susceptible host population [30]. We seek to infer the R_0 from the classical epidemic, denoted $R_0^{(1)}$ and defined by $R_0^{(1)} = \beta/\gamma_1$, and the R_0 of the new epidemic, denoted $R_0^{(2)}$ and defined by $R_0^{(2)} = \nu\beta/\gamma_2 = \nu R_0^{(1)}\gamma 1/\gamma_2$.

The doubling time of an epidemics (t_D) correspond to the time required for the number of infected hosts to double in size and it usually estimated in the early stage of an epidemics, when epidemic growth can assumed to be exponential. Here, we assumed perfect assortativity ($a_1 = a_2 = 1$) and approximated the initial exponential growth rate by $\beta - \gamma_1$ for 'classical' hosts and $\nu\beta - \gamma_2$ for 'new' hosts. Following [46], we obtain $t_D^{(1)} = \ln(2)/(\beta - \gamma_1)$ and $t_D^{(2)} = \ln(2)/(\nu\beta - \gamma_2)$.

$R_0^{(1)}$ is assumed to vary over three time intervals: from t_0 to t_1, from t_1 to t_2, from t_2 to t_f. During the first interval $[t_0, t_1]$, t_0 being the year of the origin of the epidemic in the area of Lyon, we assume that only classical hosts are present. The second interval $[t_1, t_2]$, begins in $t_1 = 1997.3$ with the introduction of the third generation HCV tests, which we assume to have decreased $R_0^{(1)}$. Finally, the 'new' hosts appear during the last interval $[t_2, t_f]$. We also wish to date the origin of this second outbreak (t_2). The final time (t_f) is given by the sampling date of our most recent sequence, which is 2018.39. The prior distributions used are summarized in Table 1.

We used a simulator implemented in `R` via the `Rcpp` package to simulate epidemiological trajectories and transmission sampled trees. The simulator resembles that developed by [29] and uses Gillespie's stochastic simulation algorithm to simulate epidemiological trajectories given our model. Further details about this simulator can be found elsewhere [preprint by Danesh et al. to be submitted to bioRxiv].

Following other phylodynamics studies, we assume that a time-scaled phylogeny of an epidemic can be correlated to a sampled transmission tree in which a branching represents a transmission event and a leaf represents a sampling event. Here, our simulator generates phylogenies of infections using the coalescent approach based on simulated trajectories and sampling dates. Importantly, we assume that the virus can still be transmitted after sampling.

We simulated $61,000$ phylogenies from known parameter sets drawn in the prior distributions

shown in Table 1. These were used to perform the rejection step and build the regression model in the Approximate Bayesian Computation inference.

ABC

Summary statistics

Phylogenies are rich objects and to compare them we used summary statistics. These were chosen to capture the epidemiological information that we wanted to extract. In particular, we used the summary statistics based on branch lengths, topology of the tree, and lineage-through-time (LTT) plot developed by [29].

We also computed additional new summary statistics to extract information regarding the heterogeneity of the population, the assortativity, and the difference between the two R_0. To do so, we annotated each internal node by associating it with a probability of being in a particular state (here the type of host, classical or new). This probability was assumed to be given by the ratio

$$P(Y) = \frac{\text{number of leaves labelled } Y}{\text{number of descendent leaves}} \tag{2}$$

where Y is a type of host.

Each node could therefore be annotated with n ratios, n being the number of possible states (i.e. types of label). Since in our case $n = 2$, we only followed one of the labels and used the mean and the variance of the distribution of the ratios (one for each node) as summary statistics.

In a phylogeny, 'cherries' are pairs of leaves that are adjacent to a common ancestor. There are $n(n+1)/2$ categories of cherries. Here, we counted the number of homogeneous cherries for each label and the number of heterogeneous cherries. Furthermore, we considered triplets, that is a cherry and a leaf adjacent to a common ancestor, and introduced three categories: homogeneous triplets, triplets whose cherries were homogeneous for a label and whose leaf was labelled with another trait, and triplets whose cherries were heterogeneous. We expected the structure of cherries and triplets capture the information about the interaction between the different hosts.

The Lineage-Through-Time (LTT) plot displays the number of lineages of a phylogeny over time: the number of lineages is incremented by one for each new branch in the phylogeny, and is decreased by one for each leaf in the phylogeny. We used the ratios defined for each internal node to build an LTT for each label type, which we refer to as an LTT label plot. After each branching event in phylogeny, we incremented the number of lineages by the value of the ratio of the internal node for the given label. This number of lineages was decreased by one for each leaf in the phylogeny. In the end, we obtained $n = 2$ LTT label plots.

Finally, for each label, we computed some of the same branch lengths summary statistics as for the unlabelled phylogeny on homogeneous clusters and heterogeneous clusters present in the phylogeny. Homogeneous clusters were defined by their root having a ratio of 1 for one type of label and their size being greater than $N_{\min}$. For heterogeneous cluster, we kept the size criterion and imposed that the ratio was smaller than 1 but greater than a threshold ϵ. After preliminary analyses, we set $N_{\min} = 4$ leaves and $\epsilon = 0.7$. We therefore obtained a set of homogeneous clusters

and a set of heterogeneous clusters the branch lengths of which were pooled into two sets to compute the summary statistics of heterogeneous and homogeneous clusters.

Regression-ABC

We first measured multicollinearity between summary statistics using variance inflation factors (VIF). Each summary statistic was kept if its VIF value was lower than 10. This step led to the selection of 88 summary statistics out of 234.

We then used the **abc** function from the `abc R` package to infer posterior distributions from rejection only. Finally, we performed linear adjustment using elastic net regression.

The abc function performs a classical one-step rejection algorithm [47] using a tolerance parameter P_δ, which represents a percentile of the simulations that are close to the target. For this, we calculate a Euclidian distance between the simulations and the target using the normalized simulated vector of summary statistics and the normalized target vector.

Prior to linear adjustment, the `abc` function performs smooth weighting using an Epanechnikov kernel [47]. Then, using the `glmnet` package in `R`, we implemented an elastic-net (EN) adjustment, which balances the Ridge and the LASSO regression penalties [48]. The EN performing a linear regression, is it not subject to the risk of over-fitting that may occur for non-linear regressions (e.g. when using neural networks, support vector machines or random forests).

We inferred posterior distributions of t_0, t_2, a_1, a_2, ν, γ_1, γ_2, $R_0^{(1),t_1}$ and $R_0^{(1),t_2}$ using our ABC-EN regression model with $P_\delta = 0.1$.

Parametric bootstrap and cross validation

Parametric bootstrap validation consisted in simulating $1,000$ transmission additional trees from parameter sets drawn in posterior distributions. We then computed summary statistics and performed a principal component analysis (PCA) on the vectors of summary statistics for the simulated and for the target data. If the posterior distribution is informative, we expect the target data to be similar to the simulated phylogenies. On the contrary, if the posterior distribution can generate phylogenies with a variety of shapes, the target data can be outside the cloud of simulated phylogenies in the PCA.

In order to assess the robustness of our ABC-EN method to infer epidemiological parameters of our BD model, we performed a 'leave-one-out' cross-validation. This consisted in inferring posterior distributions of the parameters from one simulated tree, assumed to be the target tree, using the ABC-EN method with the remaining $60,000$ simulated trees. We ran the cross-validation 100 times with 100 different target trees and measured the mean relative error of inference (MRE).

Acknowledgments

We thank Jūlija Pečerska for her help with Beast2. GD is funded by the Fondation pour la Recherche Médicale (FRM grant number ECO20170637560). GD and SA acknowledge further support from

the CNRS, the IRD and the itrop HPC (South Green Platform) at IRD montpellier, which provided HPC resources that contributed to the results reported here (https://bioinfo.ird.fr/).

References

[1] Messina JP, Humphreys I, Flaxman A, Brown A, Cooke GS, Pybus OG, et al. Global distribution and prevalence of hepatitis C virus genotypes. Hepatology. 2015;p. 77–87. Available from: https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.27259%4010.1002/%28ISSN%291527-3350%28CAT%29VirtualIssues%28VI%29HepatologyHonorRoleTopDownloadedArticles.

[2] European Union HCV Collaborators. Hepatitis C virus prevalence and level of intervention required to achieve the WHO targets for elimination in the European Union by 2030: a modelling study. Lancet Gastroenterol Hepatol. 2017;2(5):325–336.

[3] Alter MJ. Epidemiology of viral hepatitis and HIV co-infection. J Hepatol. 2006;44(S1):S6–9.

[4] Rosenthal E, Salmon-Céron D, Lewden C, Bouteloup V, Pialoux G, Bonnet F, et al. Liver-related deaths in HIV-infected patients between 1995 and 2005 in the French GERMIVIC Joint Study Group Network (Mortavic 2005 Study in collaboration with the Mortalité 2005 survey, ANRS EN19)*. HIV Medicine. 2009;10(5):282–289. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-1293.2008.00686.x.

[5] Kovari H, Ledergerber B, Cavassini M, Ambrosioni J, Bregenzer A, Stöckle M, et al. High hepatic and extrahepatic mortality and low treatment uptake in HCV-coinfected persons in the Swiss HIV cohort study between 2001 and 2013. Journal of Hepatology. 2015 Sep;63(3):573–580. Available from: http://www.sciencedirect.com/science/article/pii/S0168827815003025.

[6] Klein MB, Althoff KN, Jing Y, Lau B, Kitahata M, Lo Re V, et al. Risk of End-Stage Liver Disease in HIV-Viral Hepatitis Coinfected Persons in North America From the Early to Modern Antiretroviral Therapy Eras. Clin Infect Dis. 2016 Nov;63(9):1160–1167. Available from: https://academic.oup.com/cid/article/63/9/1160/2576619.

[7] Pradat P, Pugliese P, Poizot-Martin I, Valantin MA, Cuzin L, Reynes J, et al. Direct-acting antiviral treatment against hepatitis C virus infection in HIV-Infected patients – "En route for eradication"? Journal of Infection. 2017 Sep;75(3):234–241. Available from: http://www.sciencedirect.com/science/article/pii/S0163445317301421.

[8] Béguelin C, Suter A, Bernasconi E, Fehr J, Kovari H, Bucher HC, et al. Trends in HCV treatment uptake, efficacy and impact on liver fibrosis in the Swiss HIV Cohort Study. Liver International. 2018;38(3):424–431. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/liv.13528.

[9] Berenguer J, Jarrín I, Pérez-Latorre L, Hontañón V, Vivancos MJ, Navarro J, et al. Human Immunodeficiency Virus/Hepatits C Virus Coinfection in Spain: Elimination Is Feasible, but the Burden of Residual Cirrhosis Will Be Significant. Open Forum Infect Dis. 2018 Jan;5(1). Available from: https://academic.oup.com/ofid/article/5/1/ofx258/4804300.

[10] Boerekamps A, van den Berk GE, Lauw FN, Leyten EM, van Kasteren ME, van Eeden A, et al. Declining Hepatitis C Virus (HCV) Incidence in Dutch Human Immunodeficiency Virus-Positive Men Who Have Sex With Men After Unrestricted Access to HCV Therapy. Clin Infect Dis. 2018 Apr;66(9):1360–1365. Available from: https://academic.oup.com/cid/article/66/9/1360/4654729.

[11] van de Laar T, Pybus O, Bruisten S, Brown D, Nelson M, Bhagani S, et al. Evidence of a Large, International Network of HCV Transmission in HIV-Positive Men Who Have Sex With Men. Gastroenterology. 2009 May;136(5):1609–1617. Available from: http://www.sciencedirect.com/science/article/pii/S001650850900184X.

[12] Salazar-Vizcaya L, Kouyos RD, Zahnd C, Wandeler G, Battegay M, Darling KEA, et al. Hepatitis C virus transmission among human immunodeficiency virus-infected men who have sex with men: Modeling the effect of behavioral and treatment interventions. Hepatology. 2016;64(6):1856–1869. Available from: https://aasldpubs.onlinelibrary.wiley.com/doi/abs/10.1002/hep.28769.

[13] Pradat P, Huleux T, Raffi F, Delobel P, Valantin MA, Poizot-Martin I, et al. Incidence of new hepatitis C virus infection is still increasing in French MSM living with HIV. AIDS. 2018 May;32(8):1077. Available from: https://journals.lww.com/aidsonline/Abstract/2018/05150/Incidence_of_new_hepatitis_C_virus_infection_is.14.aspx.

[14] Ramière C, Charre C, Miailhes P, Bailly F, Radenne S, Uhres AC, et al. Patterns of Hepatitis C Virus Transmission in Human Immunodeficiency Virus (HIV)–infected and HIV-negative Men Who Have Sex With Men. Clin Infect Dis. 2019;.

[15] Virlogeux V, Zoulim F, Pugliese P, Poizot-Martin I, Valantin MA, Cuzin L, et al. Modeling HIV-HCV coinfection epidemiology in the direct-acting antiviral era: the road to elimination. BMC Medicine. 2017 Dec;15(1):217. Available from: https://doi.org/10.1186/s12916-017-0979-1.

[16] Pybus OG, Cochrane A, Holmes EC, Simmonds P. The hepatitis C virus epidemic among injecting drug users. Infection, Genetics and Evolution. 2005 Mar;5(2):131–139. Available from: http://www.sciencedirect.com/science/article/pii/S1567134804000905.

[17] Sweeting MJ, De Angelis D, Hickman M, Ades AE. Estimating hepatitis C prevalence in England and Wales by synthesizing evidence from multiple data sources. Assessing data conflict and model fit. Biostatistics. 2008 Oct;9(4):715–734. Available from: https://academic.oup.com/biostatistics/article/9/4/715/258904.

[18] Kwon JA, Iversen J, Maher L, Law MG, Wilson DP. The Impact of Needle and Syringe Programs on HIV and HCV Transmissions in Injecting Drug Users in Australia: A Model-Based Analysis. JAIDS Journal of Acquired Immune Deficiency Syndromes. 2009 Aug;51(4):462. Available from: https://journals.lww.com/jaids/Fulltext/2009/08010/The_Impact_of_Needle_and_Syringe_Programs_on_HIV.15.aspx.

[19] Pitcher AB, Borquez A, Skaathun B, Martin NK. Mathematical modeling of hepatitis c virus (HCV) prevention among people who inject drugs: A review of the literature and insights for elimination strategies. Journal of Theoretical Biology. 2018 Nov;Available from: http://www.sciencedirect.com/science/article/pii/S0022519318305666.

[20] Breban R, Arafa N, Leroy S, Mostafa A, Bakr I, Tondeur L, et al. Effect of preventive and curative interventions on hepatitis C virus transmission in Egypt (ANRS 1211): a modelling study. The Lancet Global Health. 2014 Sep;2(9):e541–e549. Available from: http://www.sciencedirect.com/science/article/pii/S2214109X14701883.

[21] Heffernan A, Cooke GS, Nayagam S, Thursz M, Hallett TB. Scaling up prevention and treatment towards the elimination of hepatitis C: a global mathematical model. The Lancet. 2019 Mar;393(10178):1319–1329. Available from: http://www.sciencedirect.com/science/article/pii/S0140673618322773.

[22] Grenfell BT, Pybus OG, Gog JR, Wood JL, Daly JM, Mumford JA, et al. Unifying the epidemiological and evolutionary dynamics of pathogens. Science. 2004;303(5656):327–32.

[23] Volz EM, Koelle K, Bedford T. Viral phylodynamics. PLoS Comput Biol. 2013;9(3):e1002947.

[24] Frost SD, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T. Eight challenges in phylodynamic inference. Epidemics. 2015;10:88–92. Available from: http://www.sciencedirect.com/science/article/pii/S1755436514000437.

[25] Pybus OG, Charleston MA, Gupta S, Rambaut A, Holmes EC, Harvey PH. The epidemic behavior of the hepatitis C virus. Science. 2001;292(5525):2323–5.

[26] Magiorkinis G, Magiorkinis E, Paraskevis D, Ho SYW, Shapiro B, Pybus OG, et al. The Global Spread of Hepatitis C Virus 1a and 1b: A Phylodynamic and Phylogeographic Analysis. PLOS Medicine. 2009 Dec;6(12):e1000198. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000198.

[27] Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). Proc Natl Acad Sci USA. 2013;110(1):228–33.

[28] Joy JB, McCloskey RM, Nguyen T, Liang RH, Khudyakov Y, Olmstead A, et al. The spread of hepatitis C virus genotype 1a in North America: a retrospective phylogenetic study. The Lancet Infectious Diseases. 2016 Jun;16(6):698–702. Available from: http://www.sciencedirect.com/science/article/pii/S1473309916001249.

[29] Saulnier E, Gascuel O, Alizon S. Inferring epidemiological parameters from phylogenies using regression-ABC: A comparative study. PLOS Computational Biology. 2017 Mar;13(3):e1005416.

[30] Anderson RM, May RM. Infectious Diseases of Humans. Dynamics and Control. Oxford: Oxford University Press; 1991.

[31] van de Laar TJW, van der Bij AK, Prins M, Bruisten SM, Brinkman K, Ruys TA, et al. Increase in HCV Incidence among Men Who Have Sex with Men in Amsterdam Most Likely Caused by Sexual Transmission. J Infect Dis. 2007 Jul;196(2):230–238. Available from: https://academic.oup.com/jid/article/196/2/230/872344.

[32] Wandeler G, Gsponer T, Bregenzer A, Günthard HF, Clerc O, Calmy A, et al. Hepatitis C Virus Infections in the Swiss HIV Cohort Study: A Rapidly Evolving Epidemic. Clin Infect Dis. 2012 Nov;55(10):1408–1416. Available from: https://academic.oup.com/cid/article/55/10/1408/323166.

[33] Panel AHG. Hepatitis C guidance: AASLD-IDSA recommendations for testing, managing, and treating adults infected with hepatitis C virus. Hepatology. 2015 Sep;62(3):932–954. Available from: http://europepmc.org/abstract/med/26111063.

[34] Romero-Severson EO, Bulla I, Leitner T. Phylogenetically resolving epidemiologic linkage. PNAS. 2016 Mar;113(10):2690–2695. Available from: https://www.pnas.org/content/113/10/2690.

[35] Worby CJ, Lipsitch M, Hanage WP. Shared Genomic Variants: Identification of Transmission Routes Using Pathogen Deep-Sequence Data. Am J Epidemiol. 2017 Nov;186(10):1209–1216. Available from: https://academic.oup.com/aje/article/186/10/1209/3860343.

[36] Wymant C, Hall M, Ratmann O, Bonsall D, Golubchik T, de Cesare M, et al. PHYLOSCANNER: Inferring Transmission from Within- and Between-Host Pathogen Genetic Diversity. Mol Biol Evol. 2018;35(3):719–733.

[37] Volz EM, Siveroni I. Bayesian phylodynamic inference with complex models. PLOS Computational Biology. 2018 Nov;14(11):e1006546. Available from: https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006546.

[38] Kühnert D, Kouyos R, Shirreff G, Pečerska J, Scherrer AU, Böni J, et al. Quantifying the fitness cost of HIV-1 drug resistance mutations through phylodynamics. PLOS Pathogens. 2018 Feb;14(2):e1006895. Available from: http://dx.plos.org/10.1371/journal.ppat.1006895.

[39] Pradat P, Caillat-Vallet E, Sahajian F, Bailly F, Excler G, Sepetjan M, et al. Prevalence of hepatitis C infection among general practice patients in the Lyon area, France. Eur J Epidemiol. 2001 Jan;17(1):47–51. Available from: https://doi.org/10.1023/A:1010902614443.

[40] D'Oliveira JA, Voirin N, Allard R, Peyramond D, Chidiac C, Touraine JL, et al. Prevalence and sexual risk of hepatitis C virus infection when human immunodeficiency virus was acquired through sexual intercourse among patients of the Lyon University Hospitals, France, 1992-2002. J Viral Hepat. 2005 May;12(3):330–332. Available from: http://europepmc.org/abstract/med/15850476.

[41] Sahajian F, Bailly F, Vanhems P, Fantino B, Vannier-Nitenberg C, Fabry J, et al. A randomized trial of viral hepatitis prevention among underprivileged people in the Lyon area of France. J Public Health (Oxf). 2011 Jun;33(2):182–192. Available from: https://academic.oup.com/jpubhealth/article/33/2/182/1586569.

[42] Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol. 2010 May;59(3):307–321. Available from: https://academic.oup.com/sysbio/article/59/3/307/1702850/New-Algorithms-and-Methods-to-Estimate-Maximum.

[43] Lefort V, Longueville JE, Gascuel O. SMS: Smart Model Selection in PhyML. Mol Biol Evol. 2017 Sep;34(9):2422–2424. Available from: https://academic.oup.com/mbe/article/34/9/2422/3788860.

[44] Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Computational Biology. 2014 Apr;10(4):e1003537. Available from: https://dx.plos.org/10.1371/journal.pcbi.1003537.

[45] Drummond AJ, Ho SYW, Phillips MJ, Rambaut A. Relaxed phylogenetics and dating with confidence. PLoS Biol. 2006;4(5):e88.

[46] Wallinga J, Lipsitch M. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc R Soc Lond B. 2007;274:599–604.

[47] Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian Computation in Population Genetics. Genetics. 2002 Dec;162(4):2025–2035. Available from: https://www.genetics.org/content/162/4/2025.

[48] Zou H, Hastie T. Regularization and Variable Selection via the Elastic Net. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2005;67(2):301–320. Available from: https://www.jstor.org/stable/3647580.